Generate rather than Retrieve: Large Language Models are Strong Context Generators

Year: 2022 | *Citations: 383* | *Authors: W. Yu, Dan Iter, Shuohang Wang*

Abstract

Knowledge-intensive tasks, such as open-domain question answering (QA), require access to a large amount of world or domain knowledge. A common approach for knowledge-intensive tasks is to employ a retrieve-then-read pipeline that first retrieves a handful of relevant contextual documents from an external corpus such as Wikipedia and then predicts an answer conditioned on the retrieved documents. In this paper, we present a novel perspective for solving knowledge-intensive tasks by replacing document retrievers with large language model generators. We call our method generate-then-read (GenRead), which first prompts a large language model to generate contextutal documents based on a given question, and then reads the generated documents to produce the final answer. Furthermore, we propose a novel clustering-based prompting method that selects distinct prompts, resulting in the generated documents that cover different perspectives, leading to better recall over acceptable answers. We conduct extensive experiments on three different knowledge-intensive tasks, including open-domain QA, fact checking, and dialogue system. Notably, GenRead achieves 71.6 and 54.4 exact match scores on TriviaQA and WebQ, significantly outperforming the state-of-the-art retrieve-then-read pipeline DPR-FiD by +4.0 and +3.9, without retrieving any documents from any external knowledge source. Lastly, we demonstrate the model performance can be further improved by combining retrieval and generation. Our code and generated documents can be found at https://github.com/wyu97/GenRead.